Explor Anim Med Res.
Vol. 14, Issue 2, 2024
DOI: 10.52635/eamr/14.2.187-200

ISSN 2277-470X (Print), ISSN 2319-247X (Online)
Website: www.animalmedicalresearch.org

Research Article

GENETIC SYMPHONY: INVESTIGATING CODON USAGE BIAS AND EVOLUTIONARY DYNAMICS IN WEST NILE VIRUS ACROSS DIVERSE GEOGRAPHICAL REGIONS

Swati Rani, Varsha Ramesh, Mallikarjun S Beelagi[1], Raaga R, Kuralayanapalya Puttahonnappa Suresh, Nagendra Nath Barman[2], Sharanagouda S Pati[1*]

Received 24 June 2024, revised 26 October 2024

ABSTRACT: West Nile Virus (WNV) infection, a significant zoonotic disease caused by *Flaviviruses*, affects birds, humans, and other wildlife species, leading to mild to severe fever and sometimes fatal neuroinvasive disease. Since its discovery, WNV has triggered epidemics on every continent except Antarctica, leading to significant financial losses from treatment costs, control programs, and the loss of animals and their products. The absence of specialized antiviral therapies or effective vaccines contributes to ongoing outbreaks in both endemic and non-endemic regions. This study combines analyses of codon usage bias, evolutionary dynamics, and phylogeography of WNV across Africa, America, Asia, and Europe. Results indicate that mutational pressure and natural selection shape codon usage bias, with natural selection being the primary driving force. Evolutionary dynamic analysis identified Uganda (Africa), China (Asia), Connecticut (United States of America), and Russia (Europe) as significant regions, with the tMRCA dated to 1908, 1785, 1852, and 1785, respectively. The highest evolutionary rate was observed in America, with a mean rate of 6.498E-2. These findings highlight key regions and evolutionary mechanisms driving WNV spread, essential for targeted surveillance and intervention. The study informs the development of specialized antiviral therapies and vaccines to mitigate WNV outbreaks globally.

Keywords: Codon usage bias, West Nile Virus, Evolutionary dynamics, Phytogeography analysis, tMRCA, Epidemiological analysis.

INTRODUCTION

West Nile virus (WNV) is a mosquito-borne zoonotic disease belonging to the *Flavivirus* genus within the *Flaviviridae* family, which includes over 75 viral species. It was initially identified in 1937 in Uganda, Africa, and has since spread globally, impacting Europe, North America, the Middle East, and West Asia [1, 2]. The virus circulates between mosquitoes and birds but can also affect mammals such as horses and humans. Since the mid-1990s, WNV outbreaks have grown in severity, frequency, and geographical range, driven by increasing prevalence of mosquito vectors [3]. After the propagation of the viral endemic across the world, it is considered the most important causative agent of viral encephalitis and a major global human health issue now.

As with other *Flaviviruses*, WNV possesses a single-stranded positive RNA genome of approximately 11kb. Its non-coding region's stem-loop structure aids in viral replication. The genome contains an open reading frame (ORF) encoding a polyprotein, further cleaved into 3 structural (C, prM/M, E) and 7 non-structural (NS1, NS2A, NS2B, NS3, NS4A, NS4B, NS5) proteins, flanked by UTRs (5' and 3') [3] (Fig. 1). The virus is encapsulated within an enveloped, icosahedral nucleocapsid and appears spherical. Genetic

ICAR-National Institute of Veterinary Epidemiology and Disease Informatics (NIVEDI), Yelahanka, Bengaluru-560064, India.
[1]Department of Biotechnology and Bioinformatics, Faculty of Life Sciences, JSS Academy of Higher Education, Mysuru-570015, India.
[2]Assam Agricultural University, Khanapara, Assam-781022, India.
[*]Corresponding author. e-mail: sharanspin13@gmail.com

variations among WNV lineages, isolates, and strains are highlighted by the core protein's 105 charged amino acid residues [4, 5].

The use of synonymous codons, which vary across genes and organisms, indicates codon usage bias (CUB) [6, 7]. This bias is influenced by genomic factors such as nucleotide composition, relative synonymous codon usage (RSCU), and the effective number of codons (ENC), which may arise from natural selection or mutational pressure [8, 9, 10]. Studying the synonymous codon usage in viruses aids in understanding their molecular evolution, particularly in light of increased prevalence and potential danger [7].

Recent research has significantly advanced our understanding of CUB in viruses. In RNA viruses, genomic nucleotide composition, particularly GC content, plays a crucial role in shaping codon usage patterns. For instance, studies have shown that host selection pressures, especially after interspecies transmission, significantly affect codon usage in viruses [11]. Additionally, CUB can impact viral protein expression and replication efficiency, as observed in various viral families like herpesvirus and lentivirus [12]. Understanding these biases not only provides insights into viral evolution and host adaptation but also offers potential strategies for antiviral development, such as manipulating codon usage to attenuate viruses for vaccine purposes.

In the realm of population genetics, discerning the diversity of a subpopulation from its ancestral groups is pivotal. Over time, random genetic mutations in ancestor populations lead to distinct differences in offspring populations, emphasizing the importance of studying population divergence within the population dynamics. Consequently, an evolutionary study involves a method of identifying the distinctly diversified sub-population from its ancestral population. The coalescent theory is an effective method for addressing genetic issues, offering both biological modeling and comprehensive statistical data. Software like BEAST, which employs the Bayesian Markov chain Monte Carlo (MCMC) techniques for phylogenetic construction, has become fundamental in evolutionary analysis and phylogenetic time tree detection [13]. BEAST allows the analysis of multiple data partitions simultaneously, facilitating single multi-locus coalescent analysis [14].

Previous research integrating CUB with the evolutionary analysis has yielded significant findings [15]. While earlier studies primarily focused on WNV in the United States [3,16], the current research investigates CUB along with evolutionary analysis, and phylogeographic patterns across the entire polyprotein of WNV from diverse ecological regions-Africa, America, Asia, and Europe. This study aims to determine survival and evolutionary evidence while comparing molecular level differences among WNV strains in each region.

MATERIALS AND METHODS

Codon usage bias analysis

Data collection and sequence alignment

The necessary data for analyzing the bias in codon usage were obtained from the NCBI (National Institute of Biological Information) virus database (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/), specifically, the coding sequences (CDS) of the WNV. The dataset was downloaded region-wise from Africa (n=168), America (n=1343), Asia (n=14), and Europe (n=157) and from three hosts specific, Humans, Horse, and Mosquito 72, 20, and 840 sequences respectively. The collected data were edited (deletion of stop codons) and aligned using the Multiple Sequence Comparison by Log-Expectation (MUSCLE) codon algorithm of MEGA-X software [17].

Nucleotide composition and dinucleotide abundance frequency

The composition of nucleotides and the content of GC across the first three positions of WNV of all the regions were calculated using the MEGA-X and R program, version 3.4.1 ("SeqinR" package) [18]. Also, abundance frequencies across all the 16 dinucleotides were calculated using the formula:

$$P_{XY} = \frac{f_{xy}}{f_y f_x}$$

In the above formula, f_{xy} is the representation of dinucleotides, and f_x represents the frequency of individual nucleotides X and Y. The >1.23 P_{XY} represents the over-representation and < 0.78 represents the underrepresented dinucleotides. Whereas, extremely over-represented and underrepresented dinucleotide frequencies were represented as $P_{XY} \geq 1.50$ and $P_{XY} \leq 0.05$, respectively. In contrast to a random assembly of mononucleotides, the XY pair is believed to exhibit either elevated or diminished relative abundance [13,15].

Relative synonymous codon usage (RSCU)

The value of relative synonymous codon usage of each codon of WNV was obtained to explore the pattern of biased usage of codons with synonymous codons [19]. The RSCU value of a codon represents the ratio of its observed frequency to its expected frequency within the synonymous codon family responsible for encoding a specific amino acid. The RSCU value of a gene or genome that obtains >1.6 is represented as over-represented, <0.6 are underrepresented codons [9].

The estimation of RSCU was done using the formula;

$$RSCUij = \frac{Xij}{\frac{1}{ni}\sum_{j=1}^{ni}(Xij)}$$

The frequency of the ith codon for the jth amino acid, which corresponds to a set of ni synonymous codons, is represented as Xij. RSCU values for WNV were computed using the R program.

Effective number of codons (ENC)

ENC serves as a metric for assessing the degree of codon usage bias in CDS, independent of gene length and the number of amino acids. When each amino acid is encoded by only one codon, indicative of an extremely biased gene or genome, the ENC range is 20-61. A lower ENC value, approaching 20, signifies high bias, while a value closer to 61 is considered unbiased. ENC values and ENC were graphically represented using the following formulas:

$$ENC = 2 + \frac{9}{F2} + \frac{1}{F3} + \frac{5}{F4} + \frac{3}{F6}$$

In the above formula, Fi in which "i = 2, 3, 4, 6", denotes an average of Fi values for i fold amino acids, where Fi can be calculated using the below formula.

$$Fi = \frac{n\sum^{i} j = 1\,(\frac{nj}{n})^2 - 1}{n - 1}$$

Wherein, 'n' represents the total sum of the observed codons for a specific amino acid, and 'nj' represents the sum of the observed jth codon for that particular amino acid.

The ENC plot was illustrated against the GC3 values of the sequence to determine the relationship between ENC and GC3 using the formula:

$$ENC^{expected} = 2 + S + \left(\frac{29}{S^2 + (1 - S)^2}\right)$$

In the above formula, S denotes the GC3 values of the sample [14].

Neutrality plot

The neutrality bias plot strategy was utilized to distinguish the components that influence the inclination of codon utilization and to inspect the extent of influence exerted by natural selection and mutational pressure among the WNV CDS. The plot determines the linear relationship between GC3 and GC12 of the sample sequence. The regression line on the neutrality plot is known as the selection-mutation equilibrium coefficient. The closer the plot slope is to zero, the less impact directional mutation pressure has on codon usage. The slope of 1 suggests that codon usage is completely neutral, with a directed mutation pressure [20, 21].

Parity rule-2 (PR2) plot

The PR2 plot was illustrated and plotted using the GC bias at the third codon position [G3/ (G3+C3)] as abscissa against AT biases at the third position [A3/ (A3+T3)]. The usage of PR2 analysis helps to identify the magnitude between natural selection and, mutational pressure. In the plot, 0.5 is the origin and interception of the X and Y-axis, and also a position where nucleotide A=T and G=C. So, when the PR2 is plotted, sequences with values at or near 0.5 indicate minimal bias [15].

Codon adaptive index (CAI)

The CAI is a method for measuring the expression gene level based on the encoding gene. The range of the CAI is 0 to 1. The highest frequent codons gain the higher adaption. In the current study, the CAI was computed utilizing DAMBE 7.0 software [22], employing a designated codon usage table as a reference for particular hosts [8, 23, 24].

Evolutionary dynamic studies

Data retrieval and sequence editing

The complete CDS of the West Nile virus were downloaded from the NCBI database regional-wise (Africa, America, Asia, and Europe), regardless of a specific host. The downloaded sequences were subjected to editing and alignment using MEGA-X software and EMBL-MAFFT. Additionally, the DAMBE tool was employed to eliminate duplicate sequences and saved the output file in Fasta and Nexus format. The algorithm GARD (Genetic Algorithm for Recombination Detection) from the Datamonkey server was applied to detect the recombination sequences in the datasets [15].

Estimation of evolutionary rate: the coalescent method

The fundamental criteria are the construction of phylogenetic analysis and the selection of statistical best-fit models for evolutionary analysis. Therefore, the phylogenetic best-fit model was determined by executing the jModel Test tool [25]. Further, the required tree operators, clock models, tree prior, and substitutional models were selected using the BEAUti program (in-built package of BEAST). Strict clock, uncorrelated relaxed clock, random local clock, and fixed local clock were used as alternative clock models with "coalescent: Bayesian Skyline tree prior" to obtain optimum results. The MCMC chain length was repetitively adjusted until all constraints reached the standard convergence criteria. The Tracer tool was then used to visualize and analyze the log files generated by BEAST, ensuring the accuracy and reliability of the estimated evolutionary rates. This detailed approach, including the careful selection and justification of models and parameters, ensures a robust and comprehensive evolutionary analysis [24].

Phylogenetic and phylogeographic study

A region-wise phylogeographic identification of WNV sequences was predicted using SPREAD software. The tree file generated after the execution of BEAST was used as an input data file. A location of virus isolation was used as a discrete location attribute to predict the rate of a lineage [26].

Determination of selection pressure (site)

The Datamonkey Adaptive Evolution (DAE) server [27] was employed to estimate the positive selection pressure from a dataset. A method for estimating selection pressure is to calculate the ratio of non-synonymous (dN) to synonymous (dS) substitution. Therefore, in the assessment of dN, dS, and dN/dS rate per site of a coding alignment sequence, the Fixed-Effects Likelihood (FEL) algorithm was employed. Also, the FEL approach assumes that the positive selection decision for each location is constant throughout the phylogeny [15].

RESULTS AND DISCUSSION

Codon usage bias analysis

Data collection and sequence alignment

The coding sequence of WNV from Africa (n=9), America (n=640), Asia (n=7), and Europe (n=157) were retrieved from the NCBI database with three hosts specific, Humans, Horse, and Mosquito 72, 20, and 840 sequences respectively, for comparison of codon adaptive index.

Nucleotide composition and dinucleotide abundance frequency

The nucleotide composition across the chosen regions of WNV was determined using MEGA-X and the R-program. The nucleotide calculation shows that WNV coding sequences are endowed with G and GC content, where components G [28.25 ± 0.13 (Africa), 28.82 ± 0.03 (America), 28.90 ± 0.28 (Asia), 28.38 ± 0.25 (Europe)], G3 [27.94 ± 0.36 (Africa), 29.51 ± 0.08 (America), 29.66 ± 0.59 (Asia), 28.26 ± 0.65 (Europe)], and GC3 [55.12 ± 0.62 (Africa), 55.76 ± 0.18 (America), 56.23± 0.94 (Asia), 55.08±0.79 (Europe)] have been utilized more often in WNV sequences all the regions (Table 1).

Also, dinucleotide abundance frequency was estimated for WNV sequences. Table 2 shows that over-represented dinucleotides as Red, underrepresented in Blue, and dinucleotides with consistent values are depicted in green color. In Africa, the observed frequencies were CA (1.25), CG (0.56), and TA (0.45). Meanwhile, in America, the represented frequencies included CA (1.31), TG (1.42), CG (0.56), and TA (0.44). Similar observations were made in the Asian region. In Europe, the observed frequencies comprised TG (1.42), TA (0.45), and CG (0.57).

The dinucleotide AA, AC, AG, CC, TC, and TT seem to match the theoretical consistent value=1. Previous WNV research [1] showed GC contents of 47.70 ± 0.21, 46.50 ± 0.11%, and 56.89 ± 0.46% for the first, second, and third positions, respectively. Additionally, research on *Flaviviruses* like WNV, dengue virus (DENV), and yellow fever virus (YFV) also highlighted GC-enriched sequences [3]. Comparing the findings to previous studies, the calculated GC contents of WNV across regions align well. Over-represented dinucleotides, CA and TG, and under-represented CG-TA were observed consistently in all regions, suggesting consistent usage of these dinucleotide sequences in WNV across diverse regions. This study affirms that GC content and specific dinucleotide frequencies are preserved across regions, suggesting a conserved codon usage pattern in WNV.

Relative synonymous codon usage (RSCU)

The synonymous usage of codons of WNV was calculated and clustered using the R-program. Fig. 2 represents the clustered heat map of WNV synonymous codon usage across the selected regions, ranging from

Table 1. WNV regional-wise nucleotide composition.

Nucleotide	Regions			
	Africa	America	Asia	Europe
T(U)	21.71± 0.08	21.65± 0.05	21.58 ± 0.10	21.71± 0.22
C	22.80 ± 0.09	22.28± 0.05	22.42 ± 0.09	22.69 ± 0.17
A	27.24 ± 0.13	27.23± 0.03	27.09 ± 0.32	27.22± 0.09
G	28.25 ± 0.13	28.82± 0.03	28.90 ± 0.28	28.38 ± 0.25
T3	20.95 ± 0.30	20.23± 0.14	20.11 ± 0.33	21.00± 0.76
C3	27.17 ± 0.30	26.25± 0.14	26.57 ± 0.38	26.82± 0.34
A3	23.91 ± 0.38	23.99± 0.08	23.64 ± 0.76	23.91± 0.12
G3	27.94 ± 0.36	29.51± 0.08	29.66 ± 0.59	28.26 ± 0.65
GC	51.05 ± 0.19	51.10± 0.07	51.32 ± 0.35	51.06± 0.30
GC1	53.56 ± 0.09	53.13± 0.07	53.34 ± 0.12	53.49± 0.14
GC2	44.49± 0.05	44.41±0.04	44.39 ± 0.07	44.60± 0.17
GC3	55.12 ± 0.62	55.76±0.18	56.23 ± 0.94	55.08± 0.79

Table 2. Regional-wise dinucleotide abundance frequency of WNV.

Dinucleotide	Dinucleotide abundance frequency, (regional-wise)			
	Africa	America	Asia	Europe
AA	1.02	0.98	1.00	1.03
AC	1.02	1.04	1.02	1.05
AG	1.00	1.01	1.00	0.97
AT	0.93	0.94	0.96	0.95
CA	1.25	1.31	1.31	0.07
CC	1.04	1.01	1.01	0.88
CG	0.56	0.56	0.55	0.57
CT	1.19	1.17	1.18	1.21
GA	1.18	1.18	1.17	1.17
GC	0.91	0.91	0.92	0.91
GG	0.99	0.99	1.01	0.95
GT	0.85	0.85	0.84	0.82
TA	0.45	0.44	0.44	0.45
TC	1.03	1.03	1.06	1.04
TG	0.81	1.42	1.42	1.43
TT	1.06	1.09	1.05	1.08

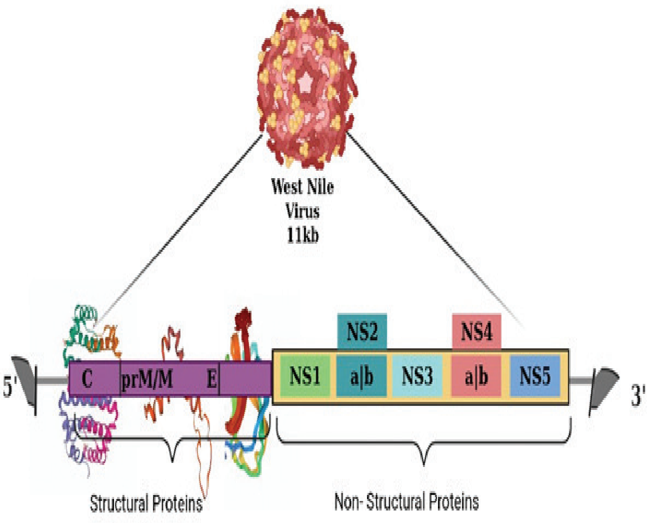

Fig. 1. Structural and non-structural proteins of WNV.

0.15 (Lowest) to 2.1 (Highest). Codons that achieve >1.6 are overrepresented and <0.6 are under-represented respectively (Fig. 3).

The RSCU of the WNV virus exhibited divergence in different regions, with 18 and 16 over-represented synonymous codons in Africa and America, and 17 over-represented codons in Asia and Europe. The over-represented codons identified among different regions' sequence commonly share most of the codons such as AAC, ACA, ACC, AGA, AGC, ATC, CCA, CTG, GCT, GGA, GTG, and TCA. The over-representation of specific codons, particularly those matching the host's tRNA pool, suggests an optimized translation efficiency, which could enhance the virus's ability to replicate rapidly within the host. This codon bias likely contributes to WNV's pathogenicity by allowing the virus to evade host immune responses more effectively. Additionally, examining the anti-codons of different hosts, such as humans, mosquitoes, and horses, provided deeper insights into how WNV adapts to diverse host species, which is crucial for understanding its transmission dynamics and potential for cross-species infection. Whereas, the previous study by Moratorio *et al.* [1] also reports that "WNV strains obtained from birds, equines, humans, and mosquitoes have almost identical synonymous codons, which are affected by relative dinucleotide frequencies".

Table 3. Substitution rate and tMRCA ages of WNV across the regions.

Regions	Substitution rate (subs/site/year)	95% HPD (Highest Posterior Density)		tMRCA age
		Low	High	
Africa	2.3311E-4	1.7899E-4	2.889E-4	1908
Asia	3.6764E-4	1.3693E-6	6.7608E-4	1785
America	6.498E-2	11.5499E-3	1.785 E-1	1852
Europe	3.4557E-4	3.0725E-4	3.8133E-4	1785

Table 4. Findings of positive selection site from WNV's coding sequences across different regions.

Regions	Positive selection site	Overall dN/dSrate ratio (ω < 1)
Africa	--	0.0289
America	2, 36, 773, 826, 900, 1262, 1367, 1839, 2209, 2364, 2377, 2522, 2950, 3388	0.0877
Asia	1415, 2294	0.0239
Europe	100, 341	0.0319

Effective number of codons (ENC)

The number of effective codon values was computed to assess the extent of the codon usage pattern in the coding sequences of WNV. The observed ENC values in WNV show slight changes across regions, averaging between 35-55 (Africa: 54.63 ± 0.45, America: 53.82 ± 0.073, Asia: 54.19 ± 0.88, Europe: 54.34 ± 0.34), indicating a moderate bias. The ENC plot (Fig. 4) suggests a predominance of mutational pressure over natural selection, which is consistent with a prior study on Flaviviruses that reported an average ENC of 55.56 [16], thus supporting our analysis of WNV. However, natural selection seems to maintain a certain degree of codon bias, which may be linked to the virus's adaptation to specific hosts.

Neutrality plot

The relationship and dominating factors (mutation pressure and natural selection) between GC12 and GC3 are investigated using a neutrality plot. A plot was generated using the GC12 and GC3 values against each other (Fig. 5).

The result shows the positive regression relationship between the GC3 and GC12 of America, Asia, and Europe as y = 0.44+0.0848, R2 = 0.16, y = 0.456+0.059, R2 = 0.67 and y = 0.47+0.0366, R2 = 0.04. In the neutrality plot, mutation pressure is considered the prevailing influence producing the bias in codon usage when the value of GC12 is statistically significantly associated with GC3 and the slope of the regression line is close to 1. In contrast, if selection is the most important factor, the regression line's slope is near 0. The current study revealed a noteworthy correlation between GC12 and GC3 across all investigated regions (Africa r = -0.687, p = 0.04, America r = 0.399, p < 0.01, Asia r = 0.821, p = 0.02, Europe r = 0.1964, p = 0.04). This serves as an indication that mutational pressure plays a predominant role in shaping the codon usage pattern in the coding sequences of WNV across all selected regions. But the calculation of regression slopes was 0.0481, 0.0848, 0.059, and 0.03696, indicating that the mutational pressure contributes 4.8-95.1%, 8.5-91.5%, 6.0-94%, and 3.7-96.3% to the mutational pressure-natural selection balance in Africa, America, Asia, and Europe respectively, among Africa, America, Asia, and Europe. Therefore, mutational pressure is taking a role but natural selection plays a more dominant factor in WNV of all different regions. This interplay between selection and mutation could be crucial in the virus's ability to adapt to varying environmental pressures and host immune systems. For instance, the observed higher contribution of natural selection in regions like Asia (94%) and Europe (96.3%) might reflect the evolutionary pressures exerted by different host populations and ecological conditions, which could influence the virulence and transmission efficiency of WNV.

Parity rule-2 (PR2) plot

The investigation into the impact of selection and mutational pressure involved the analysis of Parity Rule 2 (PR2). To illustrate the PR2 plot, the values representing AT bias in the third position were juxtaposed against GC bias in the third position. The

Table 5. Key Differences in codon usage bias and evolutionary dynamics of West Nile virus (WNV) across different regions.

Feature	Africa	America	Asia	Europe
GC3 Content (%)	55.12 ± 0.62	55.76 ± 0.18	56.23 ± 0.94	55.08 ± 0.79
Over-represented Dinucleotides	CA (1.25), TG (1.42)	CA (1.31), TG (1.42)	CA (1.31), TG (1.42)	TG (1.42), CG (0.57)
Under-represented Dinucleotides	CG (0.56), TA (0.45)	CG (0.56), TA (0.44)	CG (0.56), TA (0.44)	TA (0.45)
ENC Value Range	54.63 ± 0.45	53.82 ± 0.073	54.19 ± 0.88	54.34 ± 0.34
tMRCA	1908	1785	1852	1785
Evolutionary Rate	2.3311E-4	3.6764E-4	6.498E-2	3.4557E-4
Selection Pressure (dN/dS)	No positive site	12 positive sites	2 positive sites	2 positive sites
Primary Evolutionary Pressure	Natural selection (95.1%)	Natural selection (91.5%)	Natural selection (94%)	Natural selection (96.3%)

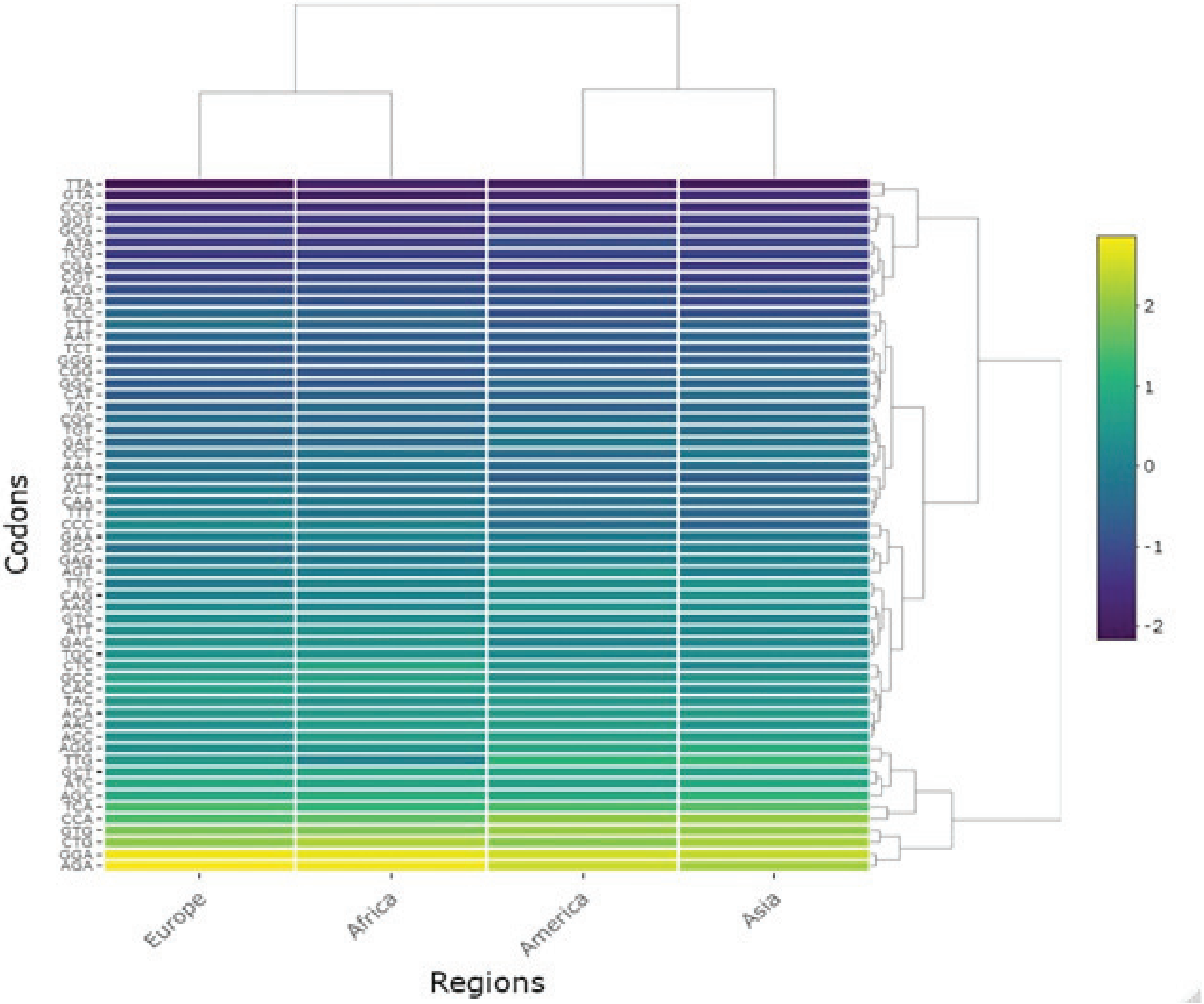

Fig. 2. The clustered heat map of RSCU for WNV sequences. [Each column indicates the regions and the row indicates a codon. The greater the RSCU value, the more abundant the codon in the sequences. The codons characterized by higher RSCU values are depicted with yellow colour and lower RSCU with blue shade].

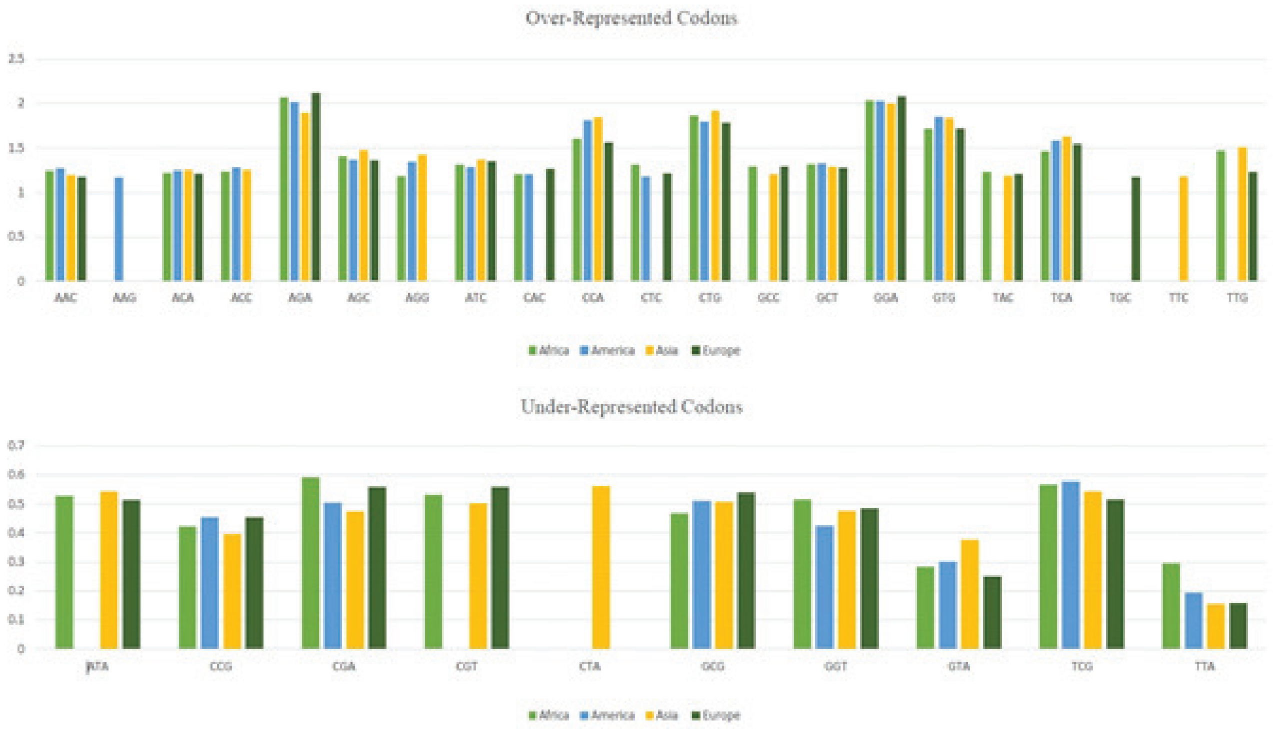

Fig. 3. Over and underrepresented codons of WNV across the selected regions' respective.

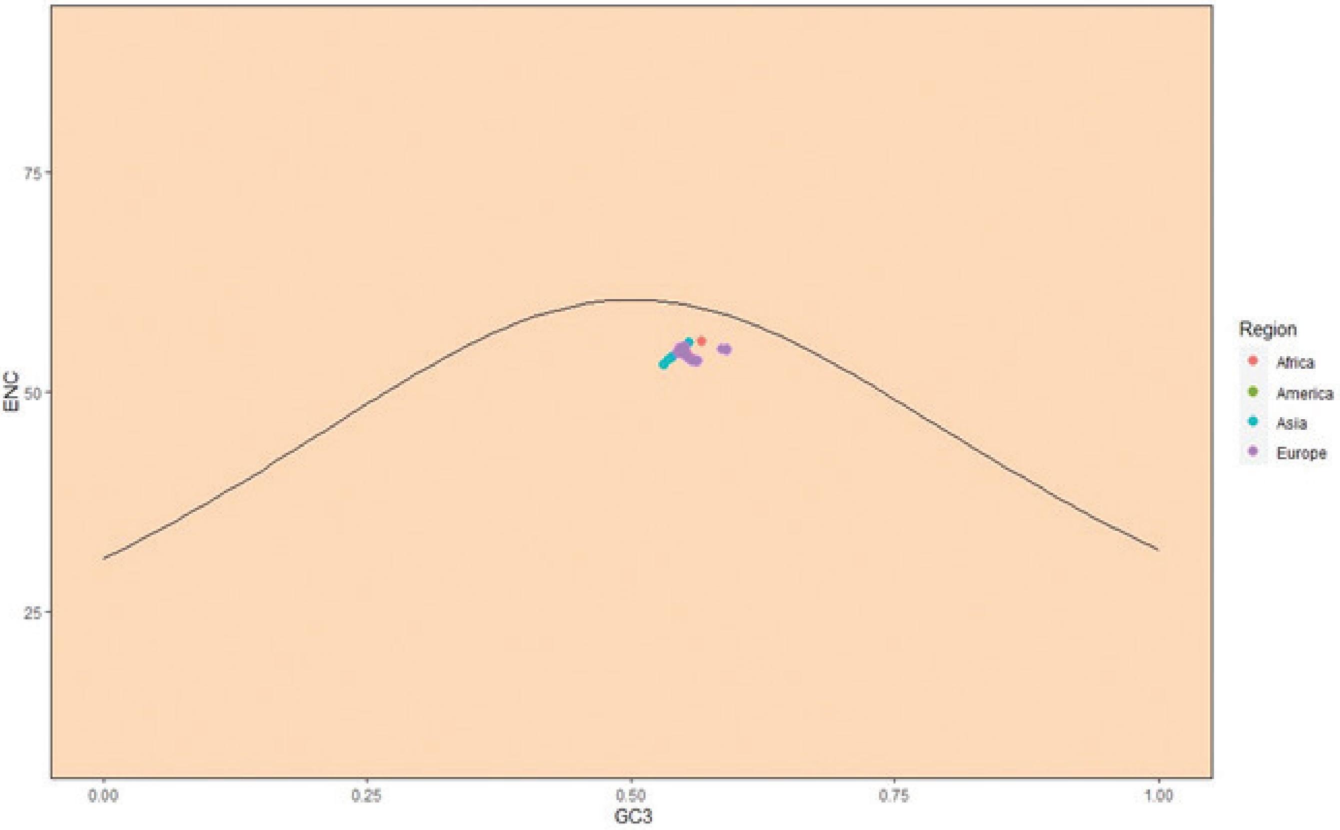

Fig. 4. Illustration of ENC-plotted against the GC3 values of WNV. The curve represents the standard expected codon usage.

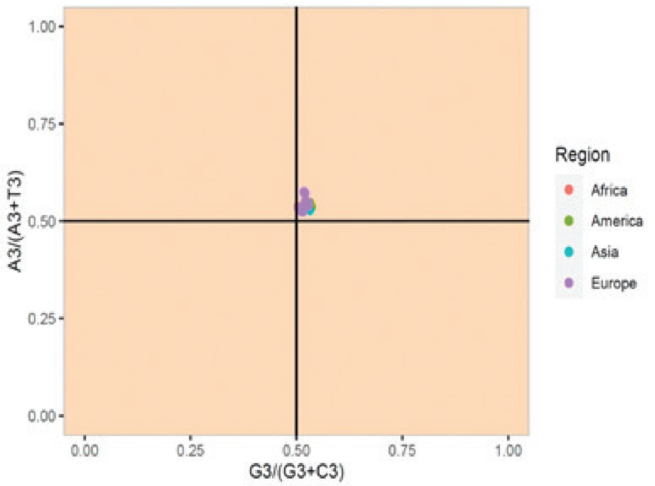

Fig. 5. Graphical representation of neutrality plot. [The regression line on a graph represents the equilibrium coefficient].

bias at [G3/ (G3+C3)] and [A3/ (A3+T3)] were plotted on the X and Y-axis, respectively. According to the methodology, if there is no bias, the nucleotide points should align with the origin of the axis (0.5). However, in this scenario, the majority of nucleotide points deviated from the origin, indicating a subtle bias in the coding sequences of WNV (Fig. 6).

Likewise, a consistent bias at the third position of GC and AT in WNV across various regions was observed. In Africa, the average bias for GC and AT was 0.50 and 0.53, respectively. In America, these values were 0.53 and 0.54, while in Asia, they were 0.53 and 0.54. In Europe, the average bias for GC and AT was 0.51 and 0.53, respectively. These consistent biases at the third position suggest a preference for purines over pyrimidines in the GC-AT composition across different regions. The strong association observed between GC3 and GC12 in the Parity Rule 2 analysis indicates that natural selection is the primary driving force, accounting for 95.1%, 91.5%, 94%, and 96.3% in Africa, America, Asia, and Europe, respectively. While natural selection plays a predominant role, mutational pressure also contributes to a slight extent, with percentages of 4.8%, 8.5%, 6%, and 3.7% in shaping the codon usage pattern across WNV in Africa, America, Asia, and Europe. The preference for purines over pyrimidines in the GC-AT composition could have implications for the stability of the viral genome and its ability to evade host immune responses. This finding is particularly relevant for vaccine development, as understanding the codon preferences of WNV can inform the design of vaccines that induce robust and long-lasting immune responses by targeting these conserved elements of the viral genome.

Codon adaptive index

The calculated CAI values, which are consistently high across all regions and hosts, indicate that WNV has a strong adaptive capacity, allowing it to efficiently replicate in diverse hosts such as humans, horses, and mosquitoes. Only WNV of human sequences were found in Africa and there were no sequences reported in horses and mosquitoes. The estimated average CAI of WNV in human host was found to be 0.79 in Africa. Whereas in America 0.79 (Humans), 0.76 (Horses), and 0.79 (Mosquitoes), Asia 0.79 (Humans), 0.77 (Horses), and 0.78 (Mosquitoes), Europe 0.79 (Humans), 0.76 (Horses), 0.78 (Mosquitoes) were identified. Thus, higher CAI values were observed consistently across all regions and hosts, indicating the higher adaptiveness of the virus across all hosts and in all selected areas.

Evolutionary dynamic studies

Data retrieval and sequence editing

Overall, 1251 WNV sequences in which, Africa (n=9), America (n=441), Asia (n=7), and Europe (n=157) were sourced from the NCBI data repository. An online server MAFFT was utilized to align the individual set of sequences. Sequence editing like removal of duplication sequence and recombination detection were evaluated using the DAMBE tool and GARD algorithm.

Estimation of evolutionary rate: the coalescent method

The optimum tMRCA was found after consecutively executing the evolutionary dynamic simulation around 1 to 20.5 million cycles for each WNV sequence set of different regions. The substitution models GTR, GTR+G (I), and HKY+G, with Coalescent: Bayesian skyline tree were employed to determine the evolutionary dynamics. As the previous study states, the genetic variation of WNV in the United States is assumed to occur in a highly defined geographic area ("evolutionary niche") where mutant virus strains accumulate genetic changes while adapting to local ecological conditions [16]. However, in the present study, as displayed in Table 3, the node age and common province of WNV from each region were determined, which revealed the substantial evolutionary rate 2.3311E-4, 3.6764E-4, 6.498E-2, and 3.4557E-4

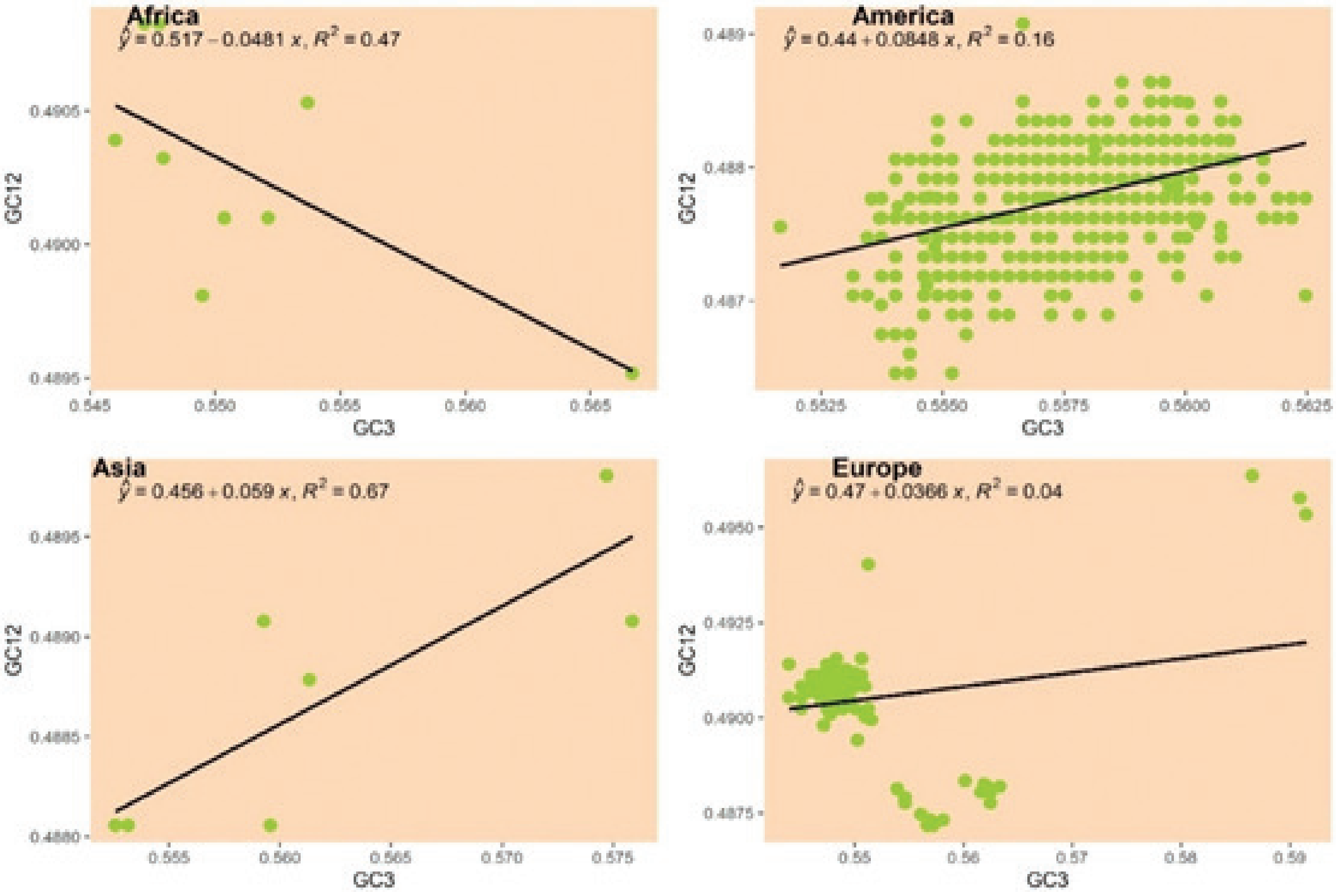

Fig. 6. PR2 plot of WNV, indicating slight bias across all the regions.

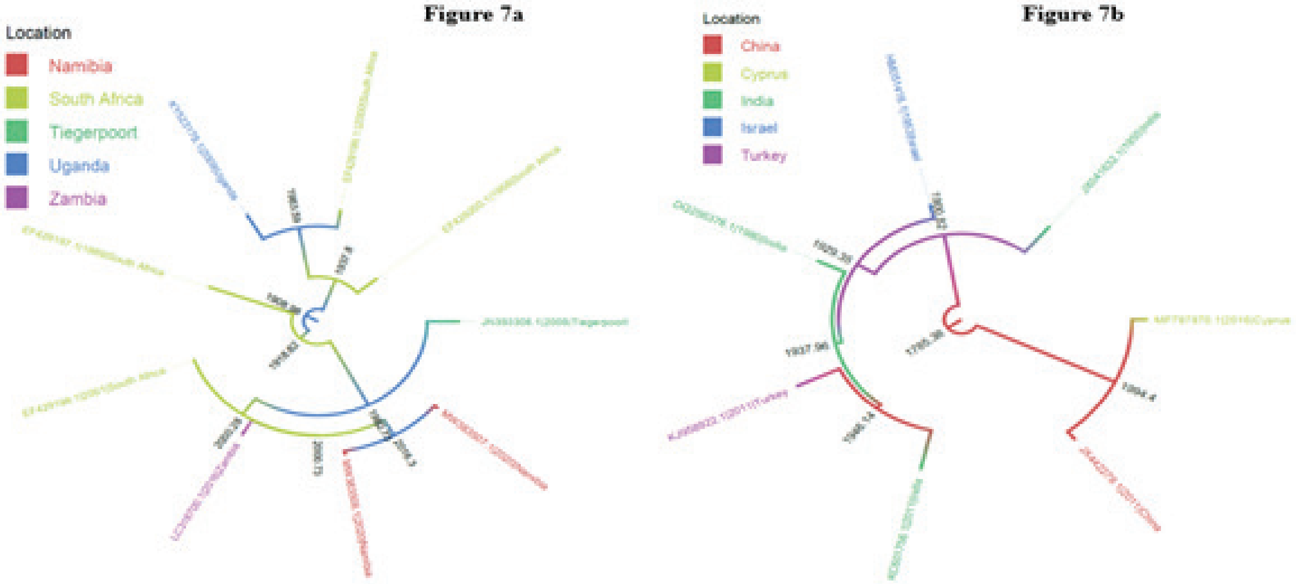

Fig. 7. Visualization of a phylogenetic tree and most common ancestor of WNV across the different regions. [(a) Africa and (b) Asia, showing tMRCA ages with a specific location].

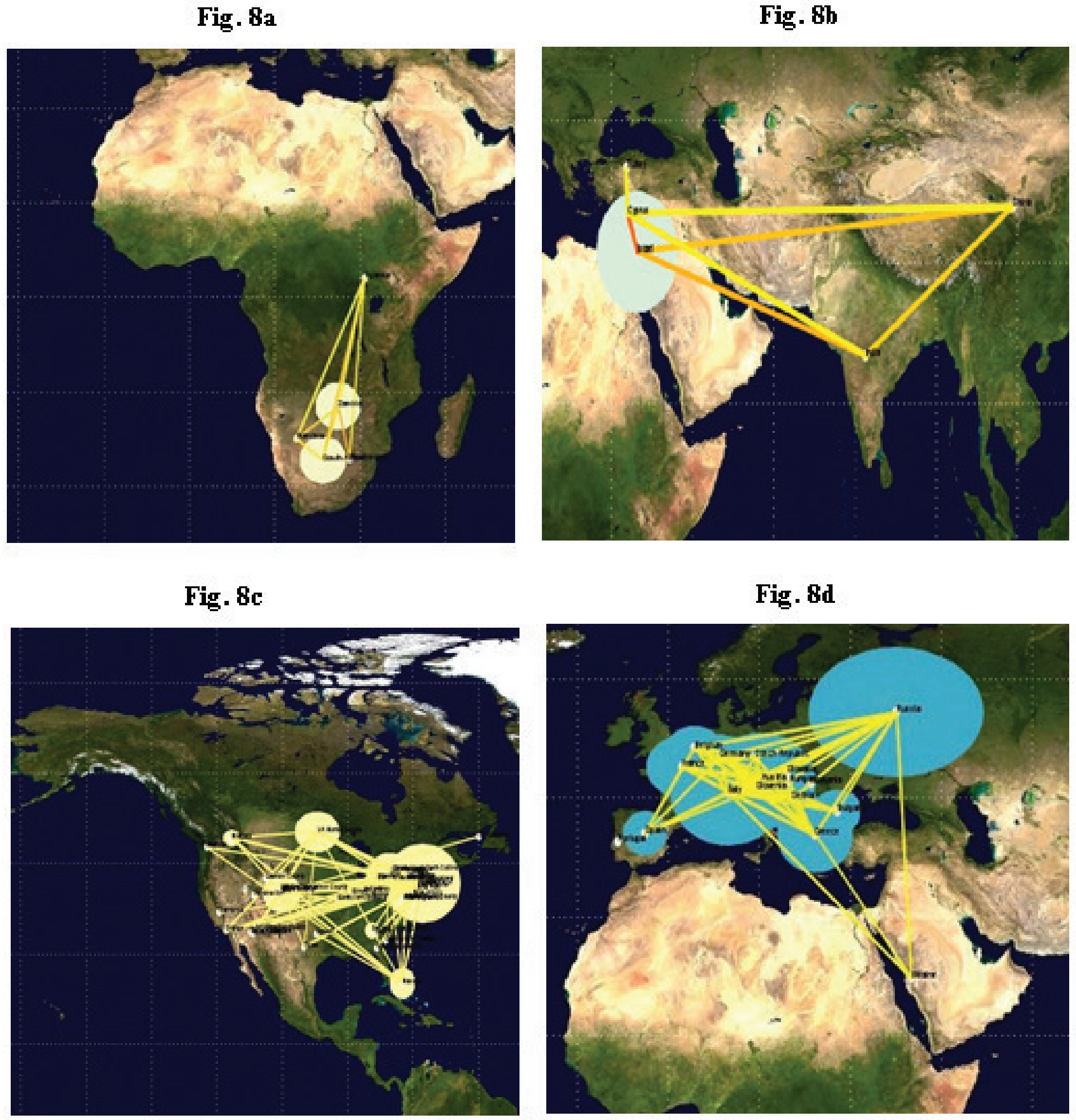
Fig. 8a

Fig. 8b

Fig. 8c

Fig. 8d

Fig. 8. The phylogeographic analysis of WNV across (a) Africa, (b) America, (c) Asia, and (d) Europe regions.

indicating the tMRCA as 1908, 1785, 1852, and 1785 of Africa, Asia, America, and Europe respectively. The older tMRCA in Africa and Asia suggests that WNV may have originated in these regions and then dispersed to other areas. Whereas, the high evolutionary rate observed in America compared to other regions suggests that WNV in America is undergoing more rapid genetic changes. This could be attributed to different ecological pressures and natural selection or virus adaptation strategies in the region.

Phylogenetic and phylogeographic study

A phylogenetic tree analysis was conducted to discern and identify the common ancestors of WNV from different regions using the tree file generated after the execution of BEAST. The Figtree and SPREAD software were used to construct the phylogenetic tree and visualize the phylogeographic lineages. The analyses reveal the evolutionary relationships and geographic distribution of WNV. Fig. 7 shows the phylogenetic analysis of WNV from each region and indicates the age and common ancestor at the root of the tree. Whereas Fig. 8 illustrates the phylogeographic status of WNV. In each region, the reddish gradient of the branch indicates the maximum lineage rate and light yellow indicates the minimum lineage rate. The circular shape on the map specifies the number of discrete states of the WNV sequences.

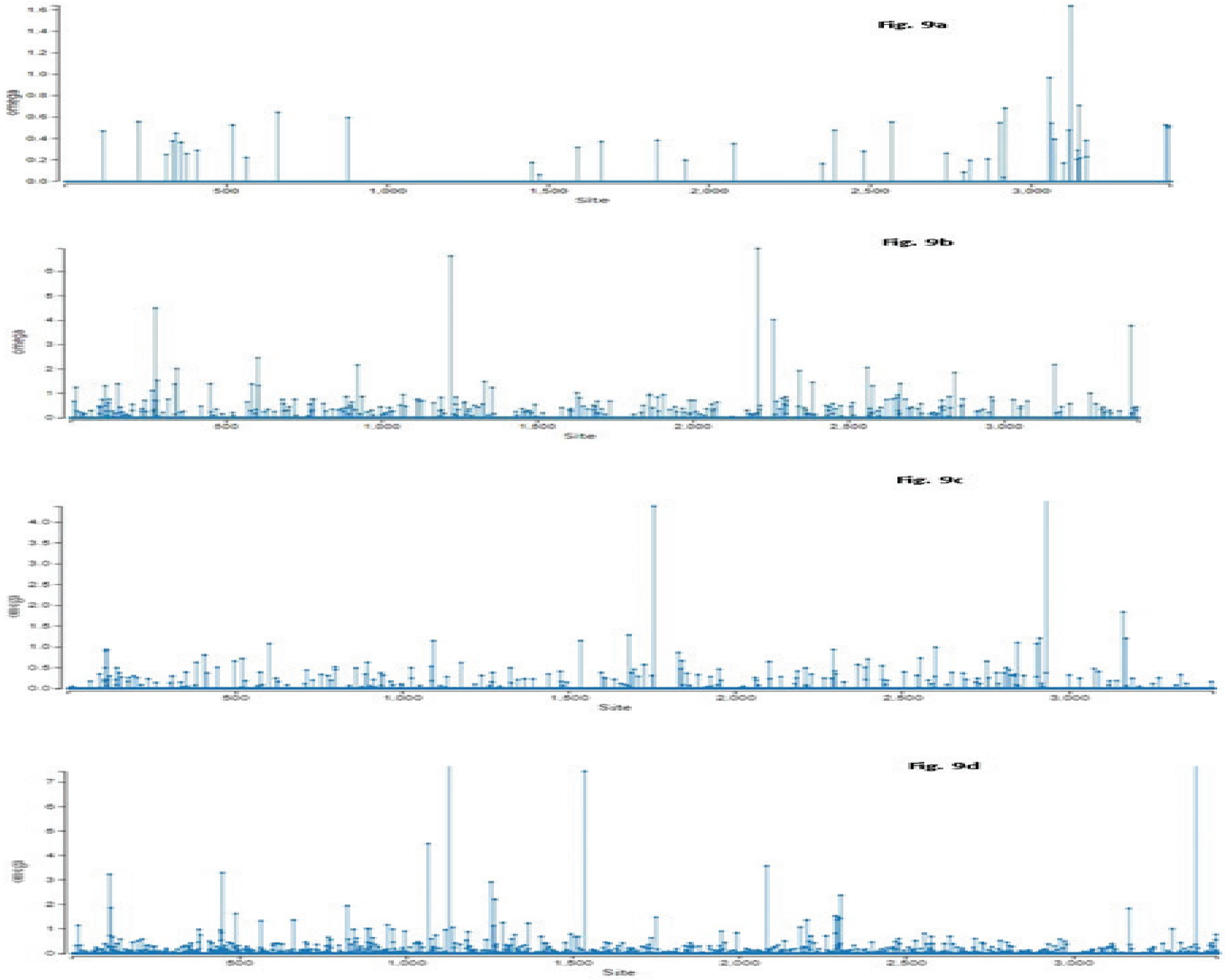

Fig. 9. Overall dN/dS (ω) ratio of WNV from different regions (a) Africa, (b) America, (c) Asia, (d) Europe.

Determination of selection pressure

The FEL algorithm from the Datamonkey server was used to assess selection pressure. The finding revealed 0,14, 02, and 08 positive sites (p-value = 0.1) with overall dN/dS rate ratios (ω) 0.0289, 0.0877, 0.0239, and 0.0319 in Africa, America, Asia, and Europe respectively (Fig. 9, Table 4). This method compares substitution rates at silent (dS) and non-silent sites (dN), with a dN/dS ratio of 1 indicating strong positive selection [15]. Taking these favorable selection characteristics we identified 12 positive selection sites in WNV from the American region, 2 sites from both Europe and Asia, and none from Africa. This aligns with the codon usage bias analysis, confirming that WNV has undergone positive selection pressure across all regions. The detection of positive selection sites in WNV, particularly in America, indicates that the virus is under strong evolutionary pressure in this region. This could be due to factors such as host immune responses or environmental conditions driving the virus to adapt rapidly. The coherence between the codon usage bias and selection pressure analyses confirms the prevalence of positive evolutionary pressures on WNV in diverse regions.

While this research has provided valuable insights into the evolutionary dynamics and codon usage bias of WNV across different regions, certain limitations should be acknowledged. The reliance on available sequence data, which may not comprehensively represent the global genetic diversity of WNV, could affect the generalizability of the findings. Additionally, the use of substitution models and coalescent methods, though effective, introduces certain assumptions and simplifications that may not fully capture the complexities of viral evolution, such as recombination and host-specific adaptation.

Moreover, while the detection of positive selection sites offers important clues about the evolutionary pressures on WNV, the analysis is sensitive to statistical thresholds, which may lead to variability in the interpretation of selection pressures. Despite these limitations, the study provides a robust framework for understanding WNV evolution, as summarised in Table 5, and these findings are a significant contribution to the field. Future research should aim to address these limitations by incorporating more comprehensive global sampling, advanced evolutionary models, and in-depth studies on host-virus interactions, which will further enhance our understanding of WNV evolution and its implications for public health.

CONCLUSION

This comprehensive study, integrating codon usage bias, evolutionary dynamics, selection pressure, and phylogeographic analyses, highlights the significant role of natural selection in shaping the codon usage patterns of West Nile Virus (WNV) across diverse regions. The findings reveal that even though natural selection act as the dominant force, varying degrees of mutational pressure also contribute to regional differences in codon usage, influenced by distinct ecological conditions. The evolutionary analysis notably identifies America as the region with the highest evolutionary rate for WNV, suggesting a more rapid epidemiologic transmission and adaptation. These insights underscore the critical need for region-specific genomic analyses to understand the molecular evolution of WNV, which could inform strategies for monitoring and controlling the virus in different ecological contexts.

ACKNOWLEDGEMENT

We express our gratitude to the Spatial Epidemiology lab at ICAR-NIVEDI for facilitating the research conducted under NICRA. Additionally, we extend our thanks to JSS AHER in Mysuru, Karnataka, India, for their essential support in facilitating this research.

REFERENCES

1. Moratorio G, IriarteA , Moreno P, Musto H, Cristina J. A detailed comparative analysis of the overall codon usage patterns in West Nile virus. Infect Genet Evol. 2013; 14:396-400, DOI:10.1016/j.meegid.2013.01.001.

2. Chancey C, Grinev A, Volkova E, Rios M. The global ecology and epidemiology of West Nile virus. Biomed Res Int. 2015; 376230, DOI:10.1155/2015/376230.

3. Behura SK, Severson DW. Bicluster pattern of codon context usages between flavivirus and vector mosquito *Aedes aegypti*: relevance to infection and transcriptional response of mosquito genes. Mol Genet Genomics. 2014; 289(5):885-894, DOI:10.1007/s00438-014-0857-x.

4. Brinton MA. The molecular biology of West Nile virus: a new invader of the western hemisphere. Annu Rev Microbiol. 2002; 56:371-402, DOI:10.1146/annurev.micro.56.012302.16065.

5. Habarugira G, Suen WW, Hobson-Peters J, Hall RA, Bielefeldt-Ohmann H. West Nile virus: An update on pathobiology, epidemiology, diagnostics, control and "one health" implications. Pathogens. 2020; 9(7):589, DOI:10.3390/pathogens9070589.

6. Ringnér M, Krogh M. Folding free energies of 5'-UTRs impact post-transcriptional regulation on a genomic scale in yeast. PLoSComput Biol. 2005; 1(7):e72, DOI:10.1371/journal.pcbi.0010072.

7. Tao P, Dai L, Luo M, Tang F, Tien P, Pan Z. Analysis of synonymous codon usage in classical swine fever virus. Virus Genes. 2009; 38(1):104-112, DOI:10.1007/s11262-008-0296-z.

8. Beelagi MS, Kumar SS, Indrabalan UB, *et al.* Synonymous codon usage pattern among the S, M, and L segments in Crimean-congo hemorrhagic fever virus. Bioinformation. 2021; 17(4):479-491, DOI:10.6026/97320630017479.

9. Deb B, Uddin A, Chakraborty S. Codon usage pattern and its influencing factors in different genomes of hepadnaviruses. Arch Virol. 2020; 165(3):557-570, DOI:10.1007/s00705-020-04533.

10. Patil SS, Indrabalan UB, Suresh KP, Shome BR. Analysis of codon usage bias of classical swine fever virus. Vet World. 2021; 14(6):1450-1458, DOI:10.14202/vetworld.2021.1450-1458.

11. Simón D, Cristina J, Musto H. An overview of dinucleotide and codon usage in all viruses. Arch Virol. 2022; 167:1443-1448, DOI:10.1007/s00705-022-05454-2.

12. Plant EP, Ye Z. Bias at the third nucleotide of codon pairs in virus and host genomes. Sci Rep. 2022; 12:4522, DOI:10.1038/s41598-022-08570-w.

13. Indrabalan UB, Suresh KP, Shivamallu C, Patil SS. An extensive evaluation of codon usage pattern and bias of structural proteins p30, p54 and p72 of the African swine fever virus (ASFV). Virusdisease. 2021; 32(4):810-822, DOI:10.1007/s13337-021-00719-x.

14. Pan S, Mou C, Wu H, Chen Z. Phylogenetic and codon usage analysis of atypical porcine pestivirus (APPV). Virulence. 2020; 11(1):916-926, DOI:10.1080/21505594.2020.1790282.

15. Beelagi MS, Indrabalan UB, Patil SS, Suresh KP, Kollur SP, *et al.* Insight of codon usage bias and evolutionary rate among the genes C, E, prM and NS5 of the Kyasanur forest disease virus. Int J Res Pharm Sci. 2021; 12(3):2028-2046, https://ijrps.com/home/article/view/176.

16. Añez G, Grinev A, Chancey C, *et al.* Evolutionary dynamics of West Nile virus in the United States, 1999-2011:

phylogeny, selection pressure and evolutionary time-scale analysis. PLoS Negl Trop Dis. 2013; 7(5):e2245, DOI:10.1371/journal.pntd.0002245.

17. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. MBE. 2018; 35:1547-1549, https://academic.oup.com/mbe/article/35/6/1547/4990887.

18. Clerc O, Frank C, Lobry JR, Penel S, Perrière G. Package 'seqinr'. 2023; https://seqinr.r-forge.r-project.org/

19. Yao H, Chen M, Tang Z. Analysis of synonymous codon usage bias in flaviviridae virus. Biomed Res Int. 2019; 5857285, DOI:10.1155/2019/5857285.

20. He B, Dong H, Jiang C. Cao F, Tao S, Xu L. Analysis of codon usage patterns in *Ginkgo biloba* reveals codon usage tendency from A/U-ending to G/C-ending. Sci Rep. 2016; 6: 35927, DOI:10.1038/srep35927.

21. Rani S, Mamathashree MN, Bharthi IU, *et al.* Comprehensive examination on codon usage bias pattern of the Bovine Ephemeral fever virus. J Biomol Struct Dyn. 2023; DOI:10.1080/07391102.2023.2258220.

22. Xia X. DAMBE7: New and improved tools for data analysis in molecular biology and evolution. MBE. 2018; 35:1550-1552, DOI:10.1093/molbev/msy073.

23. Khandia R, Singhal S, Kumar U, *et al.* Analysis of Nipah virus codon usage and adaptation to hosts. Front Microbiol. 2019; 10:886, DOI:10.3389/fmicb.2019.00886.

24. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 2007; 7:214, DOI:10.1186/1471-2148-7-214.

25. Posada D. jModel Test: Phylogenetic model averaging. Molecul Biol Evoluti. 2008; 25:1253-1256, DOI:10.1093/molbev/msn083.

26. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012; 29(8):1969-1973, DOI:10.1093/molbev/mss075.

27. Pond SL, Frost SD. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics. 2005; 21(10):2531-2533, DOI:10.1093/bioinformatics/bti320.

Cite this article as: Rani S, Ramesh V, Beelagi MS, Raaga R, Suresh KP, Barman NN, Patil SS. Genetic symphony: investigating codon usage bias and evolutionary dynamics in West Nile virus across diverse geographical regions. Explor Anim Med Res. 2024; 14(2), DOI:10.52635/eamr/14.2.187-200.